WTAC Next Generation Sequencing Course

File Formats and QC Practical Exercises

Required data

For this lab, some pre-prepared datasets are installed on the VM. Double-click on the module icon, it will open up a terminal in the directory with the data for the module.

Exercise 1: SAM header line

SAM/BAM format is the accepted standard format for storing NGS sequencing reads, base qualities, associated meta-data and alignments of the data to a reference genome. If no reference genome is available, the data can also be stored unaligned.

Download the SAM/BAM file specification document from http://samtools.github.io/hts-specs.
(direct link)

From reading page 4 of the SAM specification, look at the following line from the header of the BAM file:

```
@RG ID:ERR003612 PL:ILLUMINA LB:g1k-sc-NA20538-TOS-1 PI:2000 DS:SRP000540 SM:NA20538 CN:SC
```

Q: What does RG stand for?

Q: What is the sequencing platform?

Q: What is the sequencing center?

Q: What is the lane ID? ("Lane" is the basic independent run of a high-throughput sequencing machine. For Illumina machines, this is the physical sequencing lane. Reads from one lane are identified by the same read group ID and the information about lanes can be found in the header in lines starting with "@RG".)

Q: What is the expected fragment insert size?

Exercise 2: SAM header and samtools

Samtools comprises a set of programs for interacting with SAM/BAM files. Type `samtools` with no parameters to display the list of available commands implemented in the program. Then type `samtools view` to display a detailed usage page.

Now use the `samtools view` command to print the header of the BAM file:

```
samtools view -H NA20538.bam | less
```

Q: What version of the human assembly was used to perform the alignments? (Look for the genome assembly identifier.)

Q: How many lanes are in this BAM file? Remember that each lane is identified by a unique read group ID. Use the commands `grep` and `wc` to parse the BAM header, looking for lines starting with "@RG".

Q: What programs were used to create this BAM file?

Q: What version of bwa was used to align the reads?

Exercise 3: Alignment formats conversion

You can use samtools to convert between SAM<->BAM and to view or extract regions of a BAM file. On the command line do:

```
samtools view NA20538.bam | less -S
```

The `-S` switch causes that long lines are truncated rather than wrapped. This makes the output more readable. Alternatively, the unix command `cut` can be used to extract only the column of interest. (For example, the command `cut -f1,4` prints on output only the first and fourth columns of the input.)

Q: What is the name of the first read? (Note that the SAM specification distinguishes between the "query template", the physical sequenced molecule, and the "read", the actual sequence obtained by the experiment. However, in this exercise simply look for the QNAME field.)

Q: What position does the alignment of the read start at?

Q: What is the mapping quality of the first read?

We will convert a yeast BAM file to CRAM. In the `data` directory, there is a BAM file called `yeast.bam` that was created from *S. cerevisiae* Illumina sequencing data.

Q: Can you convert the BAM file to a CRAM file called `yeast.cram` using the `samtools view` command? First run the command without arguments to view the list of available options. For this exercise we will need `-C`, `-T` and `-o`. Note that the reference genome is stored in the file `Saccharomyces_cerevisiae.EF4.68.dna.toplevel.fa`. Name the output file `yeast.cram`.

Since CRAM files use reference based compression, we expect the CRAM file to be smaller than the BAM file. What is the size of the CRAM file?

Q: Is your CRAM file smaller than the original BAM file?

Exercise 4: VCF/BCF and bcftools

VCF/BCF format is the accepted standard format for storing variant calls with supporting data. The official specification is available from http://samtools.github.io/hts-specs.

Bcftools comprises a set of programs for interacting with VCF/BCF files. You can use bcftools to convert between VCF<->BCF and to view or extract records from a region. Using the `bcftools view` command, print the header of the BCF file:

```
bcftools view -h 1kg.bcf | less
```

Q: What version of the human assembly the coordinates refer to?

Q: Can you convert the file called `1kg.bcf` to a compressed VCF file called `1kg.vcf.gz` using the `bcftools view` command?

Similarly to BAM, VCF/BCF supports random access, that is, fast retrieval from a given region. For this, the file must be indexed.

Q: Index the BCF and extract all records from the region 20:24042765-24043073

The versatile `bcftools query` command can be used to extract any VCF field. Combined with standard UNIX commands, this gives a powerful tool for quick querying of VCFs. Try to answer the following questions with the help of the manual page.

Q: How many samples are in the BCF?
Hint: use the `-l` option.

Q: What is the genotype of the sample HG00107 at the position 20:24019472?
Hint: use the combination of `-r`, `-s`, and `-f '[ %TGT]\n'` options.

Q: How many positions there are with more than 10 alternate alleles? (See the INFO/AC tag.)
Hint: use the `-i` filtering option.

Q: List all positions where HG00107 has a non-reference genotype and the read depth is bigger than 10.

Exercise 5: Generate QC stats

We will generate QC stats for two lanes of Illumina paired-end sequencing data from yeast. We will use the `bwa` mapper to align the data to the *Saccromyces cerevisiae* genome (ftp://ftp.ensembl.org/pub/current_fasta/saccharomyces_cerevisiae/dna) and `samtools stats` to generate the stats.

Q: Read pairs are usually stored in two separate FASTQ files so that *n*-th read in the first file and the *n*-th read in the second file constitute a read pair. Can you devise a quick sanity check that reads in these two files indeed form pairs? The files must have the same number of lines and the naming of the reads usually suggests if they form a pair. The location of the files is

```
60A_Sc_DBVPG6044/lane1/s_7_1.fastq
60A_Sc_DBVPG6044/lane1/s_7_2.fastq.
```

Run the `./align.sh` script to create the mappings. The script is very short, don't be afraid to take a look inside using the command

```
less ./align.sh
```

The script contains several commands, some are combined together using pipes. (UNIX pipes is a very powerful and elegant concept which allows us to feed the output of one command into the next command and avoid writing intermediate files.)

The script will produce the BAM file `lane1.sorted.bam`. Generate the stats including only primary alignments using the command

```
samtools stats -F SECONDARY lane1.sorted.bam > lane1.sorted.bam.bchk
```

Look at the output and answer the following questions:

Q: What is the total number of reads?

Q: What proportion of the reads were mapped?

Q: How many reads were mapped to a different chromosome?

Q: What is the insert size mean and standard deviation?

Q: How many reads were paired properly? Challenge: can you verify that only mapped reads have the PROPER_PAIR bit set? (Skip the second part of this question if you don't know how to use `awk` and its bitwise `and()` operation.)

Next we will create some QC plots from the output of the `stats` command using the command `plot-bamstats` which is of the samtools package:

```
plot-bamstats -p lane1-plots/ lane1.sorted.bam.bchk
```

Now in your web browser open the html file to view the graphs `firefox lane1-plots/index.html`.

Q: How many reads have zero mapping quality?

Q: Which of the first fragments or second fragments are higher base quality on average?